Submission to The Journal of Forensic Psychiatry & Psychology

Title:

Psychological interventions for individuals with a diagnosis of Borderline Personality Disorder in forensic settings: A systematic review

Authors

Nick A. J. Stewart[a], Megan Wilkinson-Tough[a], Georgia N. Chambers[a]

[a]University of Bath, Department of Psychology, 10W, Claverton Down, Bath BA2 7AY

Corresponding author: Nick Stewart nick.stewart@bath.edu

Word count (excluding abstract, figures, references): 9087

Submission date: 1st September 2018

Revision: 17th June 2019

Declarations of interest: none

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

Thank you to Dr Cara Haines for acting as a second rater during study selection.

Abstract

Borderline Personality Disorder (BPD) is a common diagnosis in forensic settings. Associated difficulties include impulsivity and emotional dysregulation, which can create a vulnerability to impulsive acts. A BPD diagnosis is also associated with significant distress and poor physical health, making it important to understand how to tailor interventions in forensic settings. This paper reviews evidence for the effectiveness of psychological interventions designed for individuals with a diagnosis of BPD when implemented in forensic settings. A systematic search found 3485 papers, of which 13 met the inclusion criteria. The papers reported nine separate studies that implemented four distinct interventions, often adapted for particular forensic settings. Six of the nine studies included control groups. Improvements in overall 'BPD symptomatology' and specific 'BPD symptoms' were reported for all intervention types, although few differences in outcome between intervention and control groups were found. There were also reported improvements in 'BPD-related behaviours', but data on offending behaviour were absent. Heterogeneity in study quality and design makes it challenging to draw any firm conclusions about the effectiveness of any one form of intervention over another, nor about which intervention may best suit a particular setting. Research of a higher quality is needed to answer these questions.

[200 words]

Keywords: borderline personality disorder, forensic, offending, dialectical behaviour therapy, schema therapy, STEPPS

Introduction

The diagnosis of Borderline Personality Disorder (BPD) is defined as ‘a pervasive pattern of instability of personal relationships, self-image, and affects, and marked impulsivity’ (American Psychiatric Association, 2013, p. 663). The diagnosis has been critiqued for ignoring the role of childhood trauma in the aetiology of these difficulties, conceptualising them as ‘symptoms’ instead of reactions to adverse circumstances such as sexual abuse and oppression (Shaw & Proctor, 2005). In the forensic setting, BPD diagnoses are twice as common in women compared with men (Black et al, 2007), and it has been suggested (e.g., Shaw & Proctor, 2005; Wilkins & Warner, 2003) that the diagnosis has a strong relationship with gender inequality, i.e., that the abusive context in which a woman’s distress develops is ignored in favour of a label suggesting she is ‘defective’. This lack of understanding is arguably responsible for the frequent stigmatisation faced by individuals who are given the BPD label (Bonnington & Rose, 2014). Despite the problematic nature of the conceptualisations and terminology associated with the diagnosis of BPD, there is considerable consensus that individuals meeting these criteria often encounter very significant difficulties for which effective interventions are a priority. Self-injurious behaviours and suicide attempts are common, with 4% of people followed up over ten years taking their own life in one study (Zanarini et al., 2007), compared with a current 10-year suicide rate for the general population of the UK of 0.001% (Samaritans, 2016). Although the majority of people with a BPD diagnosis never commit a criminal or violent act, prevalence rates of BPD in prison populations have been found to be as high as 55% in women and 30% across genders (Black et al., 2007).

Although a limited amount of research has specifically investigated possible links between aspects of BPD and criminal acts, there is evidence to suggest that certain difficulties associated with the diagnosis could contribute to the manifestation of offending behaviour (Moore, Tull & Gratz, 2017; Herpertz, Mancke & Bertsch (2016); Raine, 1993; van den Bosch, den Haan, & Lammers, 2005, cited in van den Bosch, Hysaj, & Jacobs, 2012). Theoretical work by Linehan (1993) provides an explanation for why people with a diagnosis of BPD might be vulnerable to impulsive behaviour that could include criminal acts. Her biosocial aetiological model proposes that BPD-related problems result from a combination of emotional vulnerability in the individual and an invalidating environment early in life. The invalidating environment deprives the individual of the opportunity to learn how to understand and regulate emotions, resulting in impulsive acts when the individual feels that they are experiencing overwhelming emotional crisis (Linehan, 1993).

Another distressing experience for individuals given the diagnosis of BPD, is the frequent changeability of mood state, termed 'affect dysregulation'. This is conceptualised by Schema Theory (Young, Klosko & Weishaar, 2003), as being the result of 'flipping' between 'Schema Modes' that have developed from experiences of early adversity. Individuals are described as moving between modes representing a re-experiencing of vulnerable, angry or impulsive childhood states; modes which are introjections of harmful carer responses; and modes developed as coping responses (for example the 'bully and attack mode' is an overcompensation reaction to perceived physical and psychological threat). In this conceptualisation, offending behaviours could result from the cascade of reactions that result from the re-triggering

of early traumas (Bernstein, Arntz & Vos, 2007). Correspondingly, research into affect dysregulation has found a relationship with antisocial behaviour in studies of adolescent males and females (Mezzich et al., 1997; Snyder, 1997). A BPD diagnosis has also been found to predict institutional violence and disciplinary infractions in prison (Moore et al, 2018; Warren et al., 2002) and has been linked to an increased risk of recidivism (Black et al., 2007; Jamieson & Taylor, 2004). In correctional settings a BPD diagnosis is also associated with higher rates of other physical and psychological problems (including mood, anxiety, psychotic and eating disorders), higher suicide risk, poorer functioning and lower quality of life, in comparison with those not given a diagnosis of BPD (Black et al., 2007; Blackburn & Coid, 1999). This body of research indicates that interventions for individuals experiencing difficulties consistent with a diagnosis of BPD should be an important goal in forensic settings.

Borderline Personality Disorder has long been considered to be difficult to treat (Linehan, 1993). There is little evidence available to support the use of pharmacological interventions with such difficulties (Hancock-Johnson, Griffiths & Picchioni, 2017). However, studies on psychological interventions such as Dialectical Behaviour Therapy (DBT), Mentalization-Based Therapy (MBT) and Schema Therapy (ST) have shown that distressing experiences such as suicidal and self-destructive behaviours, anger and substance abuse are amenable to change (Bateman & Fonagy, 2008; Linehan et al., 2006; Young et al., 2003). Bloom, Woodward, Susmaras, and Pantalone (2012) conducted a systematic review of studies of DBT for BPD in inpatient settings and found that the intervention may be effective in reducing suicidal

ideation, self-injury, anxiety and depression with effect sizes ranging from very small to large. However, the authors specifically excluded any articles that related to the forensic settings and suggested that this would be a useful topic for future research.

Psychological approaches developed for typical populations cannot simply be applied to forensic populations, owing to the unique needs of forensic populations (e.g., multiple, complex and long-term mental health difficulties, frequent cognitive problems, and common experiences of severe trauma; Barnao & Ward, 2015). Considering the BPD diagnosis in particular, adaptations need to be made to take account of factors such as (a) individuals leaving custody or being transferred to another institution, which can create challenges for implementing the relatively lengthy courses of therapy usually recommended; (b) the importance of working with offence-related behaviour in addition to self-harm and suicide risk, and (c) certain challenges created by institutional environments, e.g., non-trauma informed environments and individuals living alongside others with similar difficulties. The body of evidence for interventions for people with *any* diagnosis of personality disorder in forensic settings is currently limited, which is perhaps unsurprising given the paucity of research on interventions in forensic mental health settings more generally (Barnao & Ward, 2015). Psychological interventions for individuals with a diagnosis of antisocial personality disorder (ASPD) have been reviewed systematically, with recidivism as a specific focus (Gibbon et al., 2010; Wilson, 2014). The treatment of psychopathy in forensic settings has also been subject to several reviews (e.g., Polaschek & Daly, 2013; Salekin, Worley, & Grimes, 2010). However, no systematic review has consolidated research on interventions for individuals

meeting criteria for a diagnosis of BPD in forensic settings, despite clinical advances in the area (e.g., Black, Blum, McCormick, & Allen, 2013; McCann, Ball, & Ivanoff, 2000; Nee & Farman, 2008).

This systematic review aims to identify, synthesise and critically evaluate all existing research on psychological interventions designed to help individuals with a diagnosis of BPD and its associated clinical features in forensic settings. This review is important, since national policy has for some time stipulated that services need to be improved for people with a diagnosis of BPD in forensic settings (McMurran, 2002; NIMH, 2003). The policy implementation guidance 'Personality Disorder: no longer a diagnosis of exclusion' (NIMH, 2003, p. 6) included the aim of ensuring that 'offenders with a personality disorder receive appropriate care from forensic services and interventions designed both to provide treatment and to address their offending behaviour'. Furthermore, the NICE guideline on the recognition and management of BPD is explicit that its recommendations should be applied in forensic settings (NICE, 2009). This systematic review will be highly useful for healthcare professionals and researchers working in forensic settings because it will provide them with an evidence base to justify the implementation of interventions for individuals experiencing difficulties associated with a diagnosis of BPD within their services, while also meeting national policy directives. It will also highlight gaps that should be addressed through further research.

The specific questions addressed by this review are as follows:

1. Can psychological approaches be used to treat difficulties associated with a diagnosis of BPD effectively in forensic settings?
2. Can psychological interventions developed for BPD be used effectively with individuals with other personality disorder (PD) diagnoses (i.e., PD and mixed PD) in forensic settings?
3. Are the BPD-related outcomes measured in forensic settings predominantly 'symptom-related' (e.g., emotional regulation), behaviour-related (e.g., records of incidents or challenging behaviour), or offence-related (e.g., recidivism)?

The expression of distress for individuals given a diagnosis of BPD is most commonly termed 'symptoms' in the reviewed literature and therefore this umbrella term is used to report findings to be consistent with the sources. However, the problematic nature of the terminology is recognised.

Methodology

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-analyses statement (PRISMA: Moher, Liberati, Tetzlaff, & Altman, 2009). The review protocol was published on the PROSPERO International Prospective Register of Systematic Reviews on 27 September 2016 (registration number CRD42016048373).

Inclusion Criteria

Several criteria were used to guide the selection of original research studies for inclusion in the review. The inclusion and exclusion criteria are summarised in Table 1, and justification for the key criteria is provided below.

To be considered for inclusion, studies had to have been conducted within forensic services, which were defined as those that deal exclusively with individuals who have committed offences. Such settings include both mental health and non-mental health settings, e.g., prisons, probation services, youth offender institutions/juvenile detention centres, and forensic mental health services. Studies conducted within inpatient/prison and outpatient/community settings were all considered for inclusion.

A pragmatic approach was adopted regarding BPD diagnosis within the inclusion criteria, recognising that diagnosis of personality disorders in forensic settings is variable. Although the application of reliable criteria to participant selection is important in psychological research (Spitzer, Endicott, & Robins, 1978), the authors of the present review were keen not to exclude studies that may be of value for practitioners in the field, but which might not have applied full diagnostic criteria (e.g., DSM-5) criteria for BPD to every participant. Thus, a balance was struck: to be included, all participants in a study should have a BPD or personality disorder diagnosis, either formal (i.e., confirmed by diagnostic interview) or informal (reported as a

clinical diagnosis without specific reference to a diagnostic interview). Diagnoses of 'personality disorder' or 'mixed personality disorder' were only acceptable if some participants in the study had a BPD diagnosis and provided the intervention under investigation was developed for BPD (e.g., DBT, MBT, etc.).

Initial searches indicated that few gold-standard randomised controlled trials (RCTs) would be available. It was therefore decided to include a wide range of studies that measured change on one or more measures relevant to a diagnosis of BPD following an intervention. However, studies had to have implemented distinct interventions rather than holistic service models, in which any element of the model might be responsible for therapeutic change. Service evaluation studies, risk assessment interventions and interventions aimed at staff (e.g., psychologically informed practice) were excluded. Psycho-educational interventions were only considered if changes in behaviour and/or 'symptoms' were measured (i.e., studies that solely measured changes in knowledge were excluded).

Dissertations were considered for inclusion. Papers not written in the English language were excluded since no resources for translation were available. Conference abstracts were excluded, although they were used to identify further relevant papers.

During the analysis stage, evidence from studies using different designs were considered separately from each other wherever possible. This decision recognised the greater power of controlled studies to demonstrate treatment effects in comparison with uncontrolled studies. Furthermore, randomised trials may be less susceptible to

publication biases in comparison with other study designs, since pre-specified protocols are more commonly registered for randomised trials (Reeves, Deeks, Higgins, & Wells, 2011).

[Table 1 near here]

Literature Search

The following databases were used to perform searches of titles and abstracts: PsycINFO, PsycEXTRA, MEDLINE, Embase, ProQuest Dissertations and Theses (UK & Ireland) and the International Bibliography of the Social Sciences (IBSS).

The final search was conducted on 14th August 2018. Search terms were selected to describe the setting (e.g., 'Correctional', 'Forensic', 'Prison', 'Probation'), in combination with an appropriate diagnosis (e.g., 'Borderline Personality Disorder', 'Emerging Personality Disorder', 'Personality Disorder'), in combination with an appropriate intervention (e.g., 'Cognitive Behavioral Therapy', 'Dialectical Behavior Therapy', 'Psychotherapy'). Search terms and syntax were modified to meet the requirements of the selected databases (see Appendix 1).

Selection of Studies

Titles and abstracts of all studies identified in the literature search were screened by the lead author (NS) to identify any potentially relevant studies. A subset of titles

and abstracts (10%) were assessed independently by a doctoral student (CH) to assess inter-rater agreement, which was good (kappa= 0.656). Full texts of studies that looked relevant were obtained. Reference lists of included studies, and also those of relevant existing review papers, were searched by hand to identify further relevant studies. In addition, authors of included studies were contacted via email to request further published or unpublished studies.

The full texts were then assessed by both reviewers (NS and CH) to determine eligibility for the review. Any discrepancies were resolved by discussion between reviewers. Supervision with an experienced Clinical Psychologist (MWT) was used when required. Data relating to study characteristics and outcomes were extracted by the lead author (NS) and checked by a research assistant (GC).

Assessment of Risk of Bias of Included Studies

The extent to which a review can draw conclusions about the effects of an intervention depends on the validity of the included studies (Higgins, Altman & Sterne, 2011). Therefore, risk of bias was assessed for each of the papers included within this review. This assessment was conducted using the Cochrane Collaboration's risk of bias tool, which assesses the risk of a study outcome being an underestimation or overestimation of the true effect due to certain methodological flaws (Higgins et al., 2011). Because the risk of bias tool was designed for use on randomised controlled

trials, some items (e.g., selection bias) were adapted so that they could also be applied to non-randomised and uncontrolled studies. These adaptations are described in the relevant section of the Results section.

In addition to assessing the five risk of bias items detailed in the Cochrane Collaboration's tool, i.e., selection bias (random sequence generation and allocation concealment), detection bias (blinding of outcome assessment), attrition bias and reporting bias, the researchers also assessed each study against five additional items relating to intervention integrity, described by Dane and Schneider (1998), since these were felt to be highly relevant for assessing methodological strengths and weaknesses of psychological interventions. These items were adherence bias, attention bias, programme differentiation, quality of delivery (allegiance effect) and participant responsiveness. Because no standardised guidance is available for assessing these additional sources of bias, the researchers created a document outlining agreed criteria to be used when making judgements on these items, replicating the format of the Cochrane Collaboration's tool (Higgins et al., 2011) (see Appendix 2). Reference was made to appropriate literature to define these criteria, e.g., Carroll et al. (2007), Dallimore & Griffith (2015), Dane & Schneider (1988) and Higgins & Green (2011). Risk of bias was assessed independently by one researcher (GC) and then second-scored by another researcher (NS). Any discrepancies were discussed with the lead supervisor (MWT) before the final judgements were recorded.

Results

Study Selection

The flow chart in Figure 1 shows how eligible studies were selected. The literature search a generated 2913 studies, of which 538 were identified as duplicates. After screening of titles and abstracts, 62 studies were considered eligible for full-text screening. Manual searching of reference lists of selected papers, relevant reviews, related papers and contacting researchers identified a further 12 studies. Assessment of full texts resulted in the exclusion of 61 studies: 19 were not original research (e.g. conference abstracts, reviews, opinion articles); 4 were not conducted within a forensic setting; 21 recruited participants who did not meet the BPD/PD diagnostic inclusion criteria; 4 had an inappropriate study design (e.g., qualitative, case study); 6 were excluded because the intervention did not meet the inclusion criteria (e.g., holistic service evaluations); and 7 were not written in the English language.

One study (Gee & Reed, 2013) did not include a statistical analysis of findings, and was excluded for this reason. However, the authors of the study provided a manuscript of a further unpublished study that met the inclusion criteria.

[Figure 1 near here]

Summary of Study Characteristics

Thirteen papers met the inclusion criteria (Bernstein et al., 2012; Black et al., 2008; Black et al., 2013; Black, Simsek-Duran, Blum, McCormick, & Allen, 2016; Doyle, Tarrier, Shaw, Dunn, & Dolan, 2016; Evershed et al., 2003; Gee, White, Reeves, & Bartlett, 2016; Low, Jones, Duggan, Power, & MacLeod, 2001; Nee & Farman, 2005; Nee & Farman, 2008; Santisteban et al., 2015; Tarrier et al., 2010; van den Broek, Keulen-de Vos, & Bernstein, 2011). The papers were published between 2001 and 2016. Only three of the papers were published before 2008, and eight were published since 2010. Some papers described results from the same study (i.e., preliminary results followed by either full findings or follow-up findings). Such papers were grouped together at the point of data extraction to avoid reporting the same findings twice. Thus, nine individual studies are discussed in this review. The characteristics of the nine studies are presented in Table 2.

[Table 2 near here]

Settings

Most of the studies (n=5) were conducted in the UK. Two were conducted in the USA and two were conducted in the Netherlands. Five studies (Bernstein et al., 2012; Doyle et al., 2016; Evershed et al., 2003; Low et al., 2001; van den Broek et al., 2011) were conducted in forensic/high security hospitals. Two studies (Gee et al., 2016; Nee & Farman, 2008) were conducted in prisons. One study (Santisteban

et al., 2015) was conducted in the community setting and one study (Black et al., 2013) included participants in both prison and community settings.

Participants

Sample sizes ranged from 10 participants in the smallest studies (Low et al., 2001; van den Broek et al., 2011), to 77 in the largest study (Black et al., 2013). The largest randomised controlled study (Doyle et al., 2016) included 63 participants. The mean sample size across studies was 34.11 participants. Four studies included all male participants, three studies included all female participants and two studies included a mix of genders. The mean age of participants across the nine studies was 33.11. One study (Santisteban et al., 2015) recruited adolescents, the rest included only adults.

In eight studies, personality disorder diagnoses were confirmed by clinical interview with reference to DSM-IV or DSM-III-R criteria. In the other study (Evershed et al., 2003), participants were recruited from a personality disorder service; all participants had a PD diagnosis and also met criteria for a diagnosis of BPD on the PAI (a self-report measure). Six studies included only participants with a BPD diagnosis. Three studies included participants with other PD diagnoses in addition to participants with a BPD diagnosis.

A measure of psychopathy was reported in three studies (Bernstein et al., 2012; Doyle et al., 2016; van den Broek et al., 2011), all of which used ST as an intervention. No study used psychopathy as a reason to exclude participants.

Participant drop-out rates ranged from 0% (van den Broek et al., 2011) to 52.4% (Doyle et al., 2016), with a mean of 28.92% (although it should be noted that studies varied widely in whether/how they defined drop-outs; e.g., not completing the programme, withdrawing from the study, being transferred to another site; see Table 3).

Study Design

Six studies used control groups. Three of these studies (Bernstein et al., 2012; Doyle et al., 2016; van den Broek et al., 2011) were randomised controlled trials (RCTs) versus treatment as usual (TAU), one study (Santisteban et al., 2015) was an RCT versus an active treatment, and one study (Evershed et al., 2003) was controlled (but not randomised) versus TAU. Three studies (Black et al., 2013; Gee et al., 2016; Low et al., 2001) were uncontrolled single arm studies, and one study (Nee & Farman, 2008) included both a non-randomised controlled element (12-month DBT programme) and an uncontrolled element (16-week DBT programme).

Length of intervention ranged from 16 weeks (Nee & Farman, 2008) to 36 months (Bernstein et al., 2012). Follow-up data beyond the intervention period were reported in six studies. Follow-up periods ranged from 16 weeks (Gee et al., 2016) to 52 weeks (Doyle et al., 2016). Bernstein et al. (2012) plan to collect and report three years (36 months) of follow-up data, but reported no follow-up data in their preliminary findings. The mean follow-up period for the six studies that reported follow-up data was 27.66 weeks.

Interventions

Four different forms of psychotherapeutic intervention developed for treating BPD were implemented across the studies. Four studies investigated Dialectical Behaviour Therapy (DBT; Linehan, 1993), or an adapted form of DBT. Three studies used Schema Therapy (ST; Young et al., 2003) and one study used Systems Training for Emotional Predictability and Problem Solving (STEPPS; Blum, Pfohl, John, Monahan, & Black, 2002; Blum et al., 2008). One study used Integrative Borderline Personality Disorder-Oriented Adolescent Family Therapy (I-BAFT; Santisteban, Muir, Mena, & Mitrani, 2003). Three studies offered individual therapy only, one study offered group therapy only, and five studies offered both individual and group therapy.

Dialectical Behaviour Therapy is a modified form of cognitive-behavioural therapy developed for the specific difficulties associated with a diagnosis of BPD and self-harm in the general population. The skills taught within DBT specifically target the emotional and interpersonal difficulties experienced by individuals who meet diagnostic criteria for BPD, and include approaches drawn from Eastern philosophy (Linehan, 1993). Within the selected studies, DBT was adapted for forensic settings in a range of ways, including: additional intervention targets, e.g., violent behaviour, ideation, urges and emotions (Evershed et al., 2003) or offending behaviour (Gee et al., 2016); exclusion of telephone consultation (Gee et al., 2016) or providing ward-based support in place of telephone consultation (Evershed et al., 2003); delivery of Stage 1 of DBT only (i.e., with the aim of increasing behavioural control and improving quality of life; Nee & Farman, 2008); updates to skills group materials to

make them relevant to male inpatients (e.g. adding 'watch a football match on television' to self-soothing lists; Evershed et al., 2003). Intervention length for DBT ranged from 16 weeks (Gee et al., 2016; Nee & Farman, 2008) to 12 months (Low et al., 2001; Nee & Farman, 2008) to 18 months (Evershed et al., 2003).

Schema Therapy is a form of cognitive therapy that focuses primarily on the deepest level of cognition, the Early Maladaptive Schema (EMS) and Schema Modes (Young et al., 2003). This approach was developed in response to difficulties in applying CBT when working with individuals who meet criteria for a personality disorder diagnosis, in which three common characteristics (repeated response patterns, avoidance and long-term interpersonal difficulties), together with variability of presentation, pose therapeutic challenges. Schema Therapy has been adapted for forensic settings to include a focus on specific schema modes that are hypothesised to play a role in violence and criminality (Bernstein et al., 2007). Bernstein et al. (2007) expanded the schema mode model to include modes that are more common in those who have committed offences, e.g., 'self-aggrandizer' mode and 'bully and attack' mode. Therapy in the forensic setting aims to heal an individual's vulnerable side ('vulnerable child' mode) and enhance reliance on more adaptive forms of coping ('healthy adult' mode). van den Broek et al. (2011) added arts therapy (Blacker, Watson, & Beech, 2008; Reiss, Quayle, Brett, & Meux, 1996) as an adjunctive treatment to ST.

Systems Training for Emotional Predictability and Problem Solving is a form of group therapy developed for individuals with a diagnosis of BPD that takes place

over 20 weeks (Blum et al., 2002; Blum et al., 2008). The unique element of the programme is the systems component, which includes psycho-education about the diagnosis for members of the system around an individual, encouraging them to reinforce and support the individual's new skills and manage interpersonal conflict. STEPPS was adapted for the forensic setting by Black et al. (2013) by incorporating a one-time two-hour evening event for family members and friends, corrections officers and other staff members to attend. This session included education about the BPD diagnosis and how best to respond to an individual with the disorder.

Integrative Borderline Personality Disorder-Oriented Adolescent Family Therapy was developed in recognition of the fact that behaviours associated with a diagnosis of BPD and substance use can trigger each other (Santisteban et al., 2003). The manualised intervention combines an effective intervention for adolescent substance abuse (i.e., structural family therapy) with skills components taken from DBT. Santisteban et al. (2015) implemented the programme by providing weekly family therapy with either a skills training session or an individual session each week. Individual Drug Counselling (IDC; Mercer & Woody, 1999) was used as an active comparison intervention.

Outcome Measures

A total of 47 separate psychometric measures were employed across all nine studies. These were classified into four categories for ease of reporting: BPD 'symptom-related' (i.e., relating to the diagnostic criteria for BPD), behaviour-related (i.e., behaviours that are likely to be considered problematic in forensic settings), offence-related (e.g., risk of reoffending) and mood/overall improvement (i.e., clinical measures that are not specific to a BPD diagnosis). All nine studies set out to measure changes in 'BPD symptoms' (although Bernstein et al., 2012, did not report findings within this domain in their preliminary report). Eight studies included measures of behavioural change. Three studies (Bernstein et al., 2012; Doyle et al., 2016; Nee & Farman, 2008) included offence-related measures in their design, although only Doyle et al. (2016) reported findings in this domain. Six studies also included measures relating to mood or overall improvement.

Outcomes

Table 3 provides an overview of the key findings by outcome category together with details of drop-out rates for each study.

[Table 3 near here]

BPD 'symptom-related'. Five studies included a global measure of *BPD 'symptoms'*. Four of these studies reported findings on such a measure, and all of these reported overall improvements in 'BPD symptoms'. Two of these (Nee & Farman, 2008; Santisteban et al., 2015) were controlled studies. Santisteban et al. (2015) conducted an RCT comparing I-BAFT and IDC interventions in adolescents referred by juvenile diversion programmes. Change in the constellation of difficulties associated with a BPD diagnosis was measured using the Borderline Personality Scale – Millon Adolescent Clinical Inventory (BP-MACI). 62% of participants in IDC group and 76% in the I-BAFT group were judged to have improved or recovered at 12 months (using a cut-off score of 60); however, there was no significant difference between the two intervention groups (both active). Nee and Farman (2008) conducted a non-randomised controlled trial of DBT (one-year programme) with a waitlisted control group and found a significant improvement in scores on the Borderline Syndrome Index (BSI) in the DBT group ($F(3,24)=6.98$, $p=0.002$, $ES=0.47$). However, again this change did not differ significantly from that recorded in the control group. The authors also reported an improvement on the same measure for their uncontrolled study of a 16-week DBT programme ($t(13)=2.320$, $p=0.039$).

Two uncontrolled studies also reported improvements on a *global 'BPD symptom' measure*. Black et al. (2013) implemented a 20-week STEPPS programme, at the end of which significant improvement in scores on the Borderline Evaluation of Severity Over Time scale (BEST) ($F=78.1$, $p<0.001$) was demonstrated. The large effect size ($d=1.3$) is indicative of a clinically significant change. Re-analysis of the data (Black et al., 2016) found greater improvements in BEST scores among participants who were also given an ASPD diagnosis compared with those given a BPD di-

agnosis alone (following which the authors concluded that an additional ASPD diagnosis should not be a barrier to using STEPPS). Gee et al. (2016) implemented a 16-week forensically modified DBT programme ('Options') and found an improvement in scores on the Borderline Symptom List (BSL-23), from pre- to post-treatment ($t=3.7$, $p=0.001$) with further improvement from post treatment to 32-week follow-up ($t=3.3$, $p=0.004$).

Other measures which evaluated more specific BPD-related difficulties included negative affect regulation, impulsivity, anger/irritability, dissociation, suicidality, negative cognitions/schemae and interpersonal style. Two studies reported improvements in *negative affect regulation* (Gee et al., 2016; Nee & Farman, 2008). Gee et al. (2016) found an improvement on the Negative Mood Regulation scale (NMR) from pre- to post-treatment ($t=3.9$, $p=0.001$) in their uncontrolled study, while Nee and Farman (2008) found trends towards improvement (i.e., $p<0.10$) on two of four subscales for the Emotion Control Questionnaire for both the year-long and 16-week DBT programmes; however, no between-group differences were recorded.

One study (Nee & Farman, 2008) reported significant improvements on *impulsiveness*, while two others (Doyle et al., 2016; Low et al., 2001) did not. Nee and Farman (2008) reported a reduction in impulsiveness as measured using Eysenck's Impulsivity Inventory (EII) from pre-treatment to follow-up ($F(3,24)=6.29$, $p=0.003$, $ES=0.44$) for the 12-month DBT programme and from pre-to post-treatment for their uncontrolled 16-week programme ($t(13)=3.255$ $p=0.007$).

Limited evidence was found for changes on *anger/irritability* scales. Evershed et al. (2003) found significant reductions on some subscales of the State Trait Anger Expression Inventory (STAXI) and Novaco Anger Scale (NAS) from pre-treatment to

follow-up (24 months) in their controlled DBT study, but no significant changes on other subscales. Doyle et al. (2016) did not find changes on the NAS in their RCT of ST.

Two uncontrolled studies reported reductions in *dissociative experiences* as measured on the Dissociative Experiences Scale (DES). Low et al. (2001) reported a significant reduction in dissociative experiences from pre-treatment to follow-up (i.e., 18 months; $p<0.01$) in their pilot DBT study, while Nee and Farman (2008) reported a significant pre- to post-treatment reduction during their 16-week DBT programme ($t(13)=3.363$ $p=0.006$).

Two studies (Low et al., 2001; Nee & Farman, 2008) reported some within-group improvements on measures related to *suicidality*. Low et al. (2001) found a significant reduction from pre- to post-treatment on the Beck Scale for Suicide Ideation (BSSI; $p<0.01$) and the coping beliefs subscale of the Reasons for Living Inventory (RLI), but no significant findings on the Beck Hopelessness Scale and the remaining five subscales of the RLI. Nee and Farman (2008) also found significant improvements on some subscales of the RLI but not on others.

Two studies (Doyle et al., 2016; van den Broek et al., 2011) reported findings for changes in *cognitions* that are relevant to a diagnosis of BPD. Doyle et al. (2016) reported an increase in defectiveness/shame schema in the ST+TAU group compared with TAU alone on the Young Schema Questionnaire (YSQ; estimated treatment effect at 24 months = -2.47, $p=0.008$). The authors interpret this change (which would not normally be considered a desirable outcome of ST) as potentially important in a high-risk offender population with a high prevalence of psychopathy (in which lack

of remorse and empathy are likely to be common). van den Broek et al. (2011) report a small RCT of arts therapies and ST. By measuring schema modes (similar to emotional self-states) using the Mode Observation Scale (MOS) during therapy sessions, the authors found that participants showed healthy modes significantly more frequently in arts therapy sessions than in verbal psychotherapy sessions. There was a trend towards a higher frequency of child modes in the ST condition versus TAU. The findings suggest that arts therapies and ST may be useful for evoking emotional states for individuals in forensic settings who may be difficult to reach emotionally (van den Broek et al., 2011). However, the study did not look at outcomes beyond these process factors.

Behaviour-related. Behaviour-related measures included suicidal behaviour/self-harm, aggression, disciplinary infractions, resocialisation, risk of violence and substance use. Some improvements were reported for *suicidal behaviour and self-harm*, although the only controlled study to report results in their domain (Nee & Farman, 2008) did not report a statistical analysis of findings. The authors interpreted the self-harm data for their one-year programme to indicate 'to some extent' a general downturn in the frequency of self-harm incidents for their DBT participants (although numerical data to support this claim were not reported), and there was also some evidence of a reduction in the lethality of self-harm incidents for the short-format (16-week) programme. However, the authors note that aggregate data on self-harm can be skewed by participants who experience acute self-harm episodes, an issue that is likely to be problematic for many studies conducted in forensic settings. Among the uncontrolled studies, Black et al. (2013) found that the number of suicidal and self-harm behaviours (pooled together as 'suicidal behaviour') reduced during their 20-week single-arm STEPPS programme ($t= -2.22$, $p=0.029$), while Gee et al. (2016) reported a reduction in frequency of deliberate self-harm incidents ($z= 2.9$, $p=0.003$) and number of days at active risk of self-harm and suicide ($z=3.0$, $p=0.003$) to treatment end. Low et al. (2001) report an encouraging overall trend of reductions in rates of self-harm, with an apparent post-treatment rebound effect (i.e., increased rate that subsequently decreased again). Overall, they note a reduction in self-harm in all 10 participants between pre-treatment and the final follow-up period.

Evidence of reductions in *violent behaviours* were found in one controlled and one uncontrolled trial. Evershed et al. (2003) reported a reduction in seriousness of violence-related behaviours in their DBT group versus TAU ($F=8.05$, $p=0.00$) but an equal reduction in frequency of violence-related behaviours in the two groups. Black

et al. (2008) found a reduction in disciplinary infractions (occurring in prison) from pre- to post-treatment ($t=-2.06$, $p=0.043$) in their single-arm STEPPS study. Nee and Farman (2008) recorded too few adjudications to detect a clear pattern, while Gee et al. (2016) found no significant changes in adjudications between baseline and treatment end.

Substance use (measured using both self-report and urine toxicology) decreased in both the I-BAFT and IDC intervention groups in the study by Santisteban et al. (2015), but differences in this change between the two intervention arms were not significant. Finally, Bernstein et al. (2012) examined changes in *resocialisation* (i.e., supervised and unsupervised leave) in their RCT of ST. Although the findings for the 30 participants included in the preliminary report were non-significant, the authors reported interesting trends in favour of the ST intervention: a greater proportion of participants in the ST group received both supervised and unsupervised leave compared with the TAU group, and they also received this leave more rapidly than in the TAU group. The authors interpret these findings as important clinical indications that participants in the ST group are being judged to have a lowered level of risk than TAU participants.

Offence-related. Although none of the included studies reported data on offending behaviour*, recidivism* data is being collected for participants in the large ST RCT by Bernstein et al. (2012) and will be reported when the full results are available. Nee and Farman (2008) intended to collect reconviction data, but the authors noted that these data are unlikely to become available, owing to the long sentences served by most of the participants. Two studies included measures of *offending risk*. Bernstein et al. (2012) observed a non-significant trend for scores on the Historical Clinical Risk scale (HCR-20) to improve more rapidly in those receiving ST compared to TAU in their preliminary data, while Tarrier et al. (2010) found no between-group differences on this measure.

Mood & Overall Improvement. A wide variety of measures of mood and overall improvement have been reported in studies, encompassing depression, self-esteem, locus of control, quality of life, risk (not related to offending, e.g., self-harm/suicide), personality traits, affect regulation and global therapy outcomes. Very rarely has the same outcome measure been used in two studies, making it difficult to draw broad conclusions about how effectively any particular intervention can improve these difficulties. Two studies measured *depression* (Black et al., 2013; Low et al., 2001). Both of these uncontrolled studies reported significant improvements on the Beck Depression Inventory (BDI), while Low et al. (2001) also reported an improvement on the depression subscale of the Irritability, Depression and Anxiety Scale (IDAS). Considering other mood and overall improvement measures, none of the controlled studies reported greater improvement for intervention group participants in comparison with control group participants, although Bernstein et al. (2012) report a preliminary (non-significant) finding that those receiving ST showed fewer overall negative global therapy outcomes compared to those receiving TAU over 3 years of therapy. Within-group improvements for participants receiving an intervention were found for positive and negative affect, locus of control, self-esteem and need for emergency residential treatment (Black et al., 2008; Black et al., 2016; Nee & Farman, 2008; Santisteban et al., 2015).

Assessment of risk of bias of included studies

The outcomes of the assessment of risk of bias of the included studies are summarised in Figure 2.

Selection bias

Controlled studies were judged according to whether or not a sequence generation process with adequately randomisation was used, as well as whether the allocation sequence was adequately concealed from the investigators involved in enrolling participants. 50% of the controlled studies described adequate randomisation, for example by using a remote telephone randomisation service (Tarrier et al., 2010). Two of the studies were given a high risk of bias rating as they either described participants being 'selected' (Evershed et al., 2003) or 'referred' from different establishments (Nee & Farman, 2005; 2008). One study was rated as unclear, as it had an atypical study design whereby the intervention was given to all participants regardless of allocation (van den Broek et al., 2011).

For uncontrolled studies, selection bias was judged by considering whether the researchers used a random mechanism to decide which individuals to include as participants, or whether confounding factors could have influenced who was selected and who was not. 50% of the uncontrolled studies were given a low risk of bias rating as they were secondary analyses of data routinely collected in clinical settings where the intervention is offered to all (Black et al., 2008; 2013; 2015). Two of the studies were rated as high (Gee et al., 2016; Low et al., 2001), as participants were referred by staff, which is a possible confounding factor. The remaining study (Nee & Farman, 2005; 2008) did not provide sufficient information to allow a judgement to be made.

Detection bias – blinding of outcome assessment

One of the Cochrane criteria items describes detection bias that can arise on account of inadequate blinding of participants and personnel. However, this type of blinding is unfeasible in in psychotherapy outcome research, since both participants and personnel need to be informed about the nature of the intervention being delivered in order to fully engage (Stoffers et al., 2012). Therefore this item was not assessed.

Detection bias was judged according to whether or not those assessing participant outcomes were blinded to which intervention the participant had received, or, in uncontrolled studies, whether a participant had received an intervention at all. Twenty-five percent of studies were rated as having a low risk of detection bias, two of which because the researchers described adequate blinding of outcome assessors and another one because there was good agreement between the outcome assessors and blinded second-scorers. One study was judged to have a high risk of detection bias (Santisteban et al., 2015), as it was explicitly stated that outcome assessors were not blinded to which intervention each participant had received. The remaining two-thirds of the studies, including all of the uncontrolled studies, received unclear ratings, as they did not give sufficient information about outcome assessment to allow a judgement to be made.

Attrition bias

Judgements regarding attrition bias were based on whether, firstly, there was any missing outcome data, and, secondly, whether the reason for any missing outcome data was likely to be related to the intervention outcome. The majority of studies

(67% of controlled studies and 83% of uncontrolled studies) received an unclear rating. In most cases, this was primarily because insufficient information was provided about the reasons for participant drop-out; therefore, it is possible that participants dropped out because the intervention was ineffective, which would bias the estimation of the effect in favour of the intervention. Two controlled studies (Santisteban et al., 2015 and van den Broek et al., 2011) received a low rating; the former because there were similar numbers of drop-outs in each group, and similar reasons for drop-out, and the latter because there were no missing outcome data. One uncontrolled study (Black et al., 2008) also received a low rating as there was a low drop-out rate and the reasons recorded for drop-out were not related to the intervention outcome.

Reporting bias

An assessment of reporting bias was made based on whether there was evidence of selective outcome reporting, evidenced either by inconsistency between a study protocol and the outcomes reported in the published paper, or through omission of expected outcome variables in the results. One quarter of the studies were judged to have a high risk of reporting bias. Although no protocol was found for these studies (Bernstein et al., 2012; Nee & Farman, 2005; 2013, both one-year and short programmes), outcomes were mentioned in the method section which were then not fully reported in the results. Only one study (Tarrier et al., 2010) received a low rating, as this was the only study with a protocol (published retrospectively) and all of the pre-specified measures were reported. The remaining two thirds of studies received an unclear rating: although all outcome measures specified in the method were reported in the results, either no study protocol existed or none could be found.

Intervention integrity

Adherence

Only one study (Bernstein et al., 2012) detailed an objective method of assessing adherence to the intervention protocol, which appeared to be good. Half of the controlled studies (Tarrier et al., 2010, van den Broek et al., 2011 and Nee & Farman, 2005; 2013) and one of the uncontrolled studies received high ratings (Nee & Farman, 2005; 2013); for the former two this was because although an objective method of assessing adherence was specified, adherence was found to be poor, and for the latter, because it was suggested that adherence was poor. The remaining studies were rated as unclear, either because there was no method of assessing adherence, or the reported method was deemed to be insufficiently objective, or because the outcome of the assessment of adherence was not reported.

Attention bias

Of the controlled studies, all but one (i.e., n=5) were rated as having a high risk of attention bias, because an unequal amount of attention was provided to each treatment group, and no attempt was made to control for this in data analysis. In 4 of these studies, more attention was given to the intervention group than to the control group. In only one study (Santisteban et al., 2015) was an equal amount of attention given to each treatment group, justifying a low risk of bias rating.

Programme differentiation

In five of the 12 studies, participants continued to receive interventions additional to the intervention under investigation during the study period. Because the extent of this was not measured or controlled for, these studies received a high risk of bias rating. Only one study (Black et al., 2008) received a low risk of bias rating, having specified that participants did not receive any psychosocial interventions other than the intervention under investigation. The remaining 50% of studies did not provide sufficient information for a judgement to be made, resulting in unclear ratings.

Allegiance bias

Fifty per cent of the studies were judged to have a high risk of allegiance bias because one or more of the researchers either developed or made a significant adaptation to the intervention, and the potential influence of this was not considered. The other 50% of studies were conducted by researchers who were understood not to have developed the intervention, justifying a low risk of bias rating. Notably, none of the studies considered the impact of implementer enthusiasm, whether conducted by researchers who had developed the intervention or not.

Participant responsiveness

Three-quarters of the studies included no mention of formal measures of responsiveness, enthusiasm, participation or satisfaction, and were therefore given unclear ratings. The remaining quarter of studies (Black et al., 2008; 2013; 2015) received a low risk of bias rating, as formal measures of attendance and satisfaction indicated positive levels of participant responsiveness in each of these uncontrolled studies.

[Figure 2*a* near here]

[Figure *2b* near here]

[Figure *2c* near here]

[Figure *2d* near here]

Discussion

This systematic review aimed to synthesise existing research on psychological interventions designed to help individuals with a diagnosis of BPD in forensic settings. The Borderline Personality Disorder diagnosis is associated with traits such as impulsivity and emotional dysregulation that can make individuals who meet the diagnostic criteria vulnerable to carrying out impulsive acts, including offending behaviour. Furthermore, in the forensic population, features associated with a BPD diagnosis increase the likelihood of a range of physical and psychological problems, increased suicide risk, compromised functioning and poorer quality of life. Therefore, such difficulties form an important target for interventions in institutions that aim to rehabilitate those who have committed offences. The current review aimed to provide healthcare professionals and researchers with an evidence base to justify the

implementation of interventions for this service user group within their services, and an indication of the gaps that should be addressed through further research.

Despite the relatively large number of papers identified through initial searches, a relatively small number of these met the inclusion criteria. Many studies (n=21) were excluded for the absence of appropriate participant diagnosis. Arguably, we could have used less stringent inclusion criteria with regard to diagnosis, and such an approach would perhaps better align with the reality of forensic settings where individual formulation is often considered a more informative approach than diagnosis for working with individuals with complex difficulties. However, in the context of research, diagnosis has been argued to be important for judging whether a study's findings are valid for the population in question and enabling the comparison of different study outcomes. Thus, the absence of appropriate participant diagnosis may be considered a shortcoming of existing research in this area. Researchers attempting to address this problem in the future will need to grapple with an additional question: whether to use the categorical personality disorders definitions in DSM-5, the newer trait-based systems offered by the Alternative Model for Personality Disorders in DSM-5 (see Oldham, 2015) and the new ICD-11 classification system (see Bach & First, 2018). Alternatively, research that looks to ways of systematically employing non-diagnostic, formulation-based approaches (e.g., Johnstone & Boyle, 2018) could provide a greater depth of understanding that would not divorce the manifestation of ways of coping from the early adversity experienced by individuals. A consensus around the categorisation of difficulties is needed for clearer conclu-

sions to be drawn; however, these newer classification systems allow greater complexity to be captured; this may prove helpful for assessing how well psychological therapies perform within meaningful subgroups.

The papers that did meet the inclusion criteria demonstrate how a broad range of mainstream interventions developed for individuals with a diagnosis of BPD have been adapted creatively to meet the specific demands of forensic settings. Encouraging improvements have been reported across a wide range of clinically relevant outcomes, and on some forensically relevant outcomes. Improvements have been reported on global measures of 'BPD symptomatology', specific 'BPD-related symptoms', mood and other indicators of positive mental health and functioning. In addition, reductions in harmful behaviours such as suicide and self-harm have also been noted. Importantly, no intervention appears to cause participants harm.

However, the available research must be viewed in the context of its limitations. Existing studies are variable in quality and design, and most have relatively small sample sizes. Furthermore, there has been little consistency in the types of outcomes measured, or in the specific measures employed. Consequently, it is challenging to synthesise findings to date, and meta-analysis is impossible. Based on this limited evidence is not yet possible to draw conclusions about which types of psychological intervention designed for individuals with a diagnosis of BPD may lead to a better outcome than any other, nor to recommend particular types of treatment for particular forensic settings.

A major limitation of the available evidence is the lack of well-designed controlled studies. Difficulties associated with a diagnosis of BPD, especially self-injury and suicidal attempts, tend to improve over time (Zanarini, Frankenburg, Hennen, & Silk, 2003), so evidence of improvements in uncontrolled studies must be interpreted with caution. Although most studies (n=6) reviewed here used control groups, only four of these were RCTs, the 'gold standard' design in outcomes research. One of these studies (Santisteban et al., 2015) compared two active forms of treatment, making it impossible to say what advantage either intervention may have had versus no treatment, or versus TAU. In many cases, while controlled studies found treatment effects over time (i.e., within-group differences), they failed to find differences in outcome between active treatments and TAU. Where improvements are observed in uncontrolled single arm studies, it is impossible to exclude the possibility that the improvements observed might have occurred in the absence of a BPD-specific intervention.

This review has revealed several potential sources of bias in existing research in this area. These include adherence bias, attention bias, programme differentiation bias and allegiance bias, each of which was especially evident in the controlled studies. In many cases, published papers include insufficient information to enable researchers to judge whether bias could have been introduced, and addressing this issue would help to enable greater confidence in the quality of research.

Having a BPD diagnosis has been associated with an increased risk of reoffending (Black et al., 2007; Jamieson & Taylor, 2004). Yet only a small number of studies

set out to measure recidivism and none of these studies has yet reported findings on this outcome. This presumably relates to the practical challenges of measuring recidivism (e.g., Nee & Farman, 2008). However, researchers in the field will need to make efforts to measure recidivism as an outcome if a key question for this field can be answered, i.e., can treating difficulties associated with a diagnosis of BPD reduce reoffending behaviour? That said, it is encouraging that most studies included behavioural measures in addition to psychometric measures, since these are likely to be the most meaningful outcomes to clinicians working in this field and could represent a reduction in offence-related behaviours.

Research Implications

Although interest in this area is not new, the evidence base for interventions to help individuals experiencing difficulties associated with a diagnosis of BPD in forensic settings is still in its infancy. Results to date are sufficiently encouraging to merit testing on larger scale with more rigorous designs. New studies should make efforts to focus on a smaller number of meaningful outcomes, i.e., those that measure change in core features described by the BPD diagnostic criteria (e.g., the BEST, for which sensitivity to clinical change over time has been demonstrated; Pfohl, 2009), together with behaviour- and offence-related outcomes that are relevant to the problems encountered by those with BPD-related difficulties who have committed offences. Efforts should be made to align the specific measures employed, both for characteristic difficulties (e.g. the BEST for experiences related to BPD diagnostic criteria) and for more general distress (e.g., the BDI for depression) so that they can then be more readily compared. This would make meta-analysis of outcomes possible in the future.

Researchers in the field have been creative in adapting interventions such as DBT for forensic settings. A disadvantage of this approach is that it makes it more difficult to compare different studies that use variations of the same approach (e.g., DBT delivered in its 'pure' 12-month form versus 16-week programmes with adaptations). One way to mitigate against this problem would be for researchers to share treatment protocols to encourage adapted programmes to be replicated elsewhere. It would also be helpful for researchers to pay greater attention to therapeutic process effects in addition to overall outcomes, e.g., by taking measures weekly rather than only pre- and post-intervention, and by complementing traditional RCTs with single-case series designs. Such approaches would also be helpful for understanding, for example, how length of therapy affects outcomes, and the relative contributions of individual or group therapy to outcomes.

The relatively high drop-out rates recorded in many studies reflects a formidable challenge for research in forensic settings. A frequent reason reported for this phenomenon is the transfer of participants to other sites. This poses a challenge not only for research, but for clinical intervention since standard protocols for interventions such as DBT take up to two years to implement. The promising results reported for shorter-term modifications of these interventions (Black et al., 2013; Gee et al., 2016; Nee & Farman, 2008) should provide impetus for further controlled studies of these adaptations.

Strengths and Limitations of These Findings

The search strategy used for this review was rigorous and comprehensive. The likelihood of publication bias was reduced by considering dissertations and conference abstracts and by hand searching reference lists and contacting authors to find out about unpublished manuscripts. Reliability checks and independent screening by two independent researchers were employed to ensure that the inclusion and exclusion criteria were applied with rigour. The risk of bias analysis will help individuals working in this field to judge the validity of the evidence currently available in this field.

There were several limitations. Foreign language studies were excluded owing to the absence of resources for translation. It is therefore possible that relevant studies have been excluded. In addition, synthesis of the results relied upon qualitative analysis, since meta-analysis was precluded by the heterogeneity of studies included. The relatively strict inclusion criterion around the diagnosis of participants led to the exclusion of studies that may have utility for practitioners in the field. The inclusion criteria privileged studies that used clinical diagnosis to select participants. It could be argued that studies that recruit based on partial diagnosis or specific forms of distress (e.g., McCann et al., 2000) could better reflect clinical reality.

Conclusion

Research investigating whether difficulties associated with a BPD diagnosis can be treated effectively in forensic settings has yielded promising findings. Clinicians have adapted a range of interventions creatively across a breadth of forensic settings, resulting in positive changes across a range of relevant difficulties and behaviours. However, a limited body of research, design limitations and a lack of reported between-group differences together make it difficult to assess what benefits may be afforded by treatments specifically designed for a BPD diagnosis over non-specific forms of intervention. It is also currently not possible to recommend any specific treatments over any other, nor to recommend a specific treatment for a particular forensic setting. It is hoped that this review will provide impetus and ideas for researchers in the field to add to the available evidence base, and so enable firmer conclusions to be drawn.

References

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th Ed.). Arlington, VA: American Psychiatric Publishing.

Bach, B., & First, M. B. (2018). Application of the ICD-11 classification of personality disorders. *BMC Psychiatry*, 18(1), 351.

Barnao, M., & Ward, T. (2015). Sailing uncharted seas without a compass: A review of interventions in forensic mental health. *Aggression and Violent Behavior, 22,* 77-86. doi: 10.1016/j.avb.2015.04.009

Bateman, A., & Fonagy, P. (2008). 8-year follow-up of patients treated for borderline personality disorder: Mentalization-based treatment versus treatment as usual. *The American Journal of Psychiatry, 165*(5), 631. doi: 10.1176/appi.ajp.2007.07040636

Bernstein, D. P., Arntz, A., & Vos, M. D. (2007). Schema Focused Therapy in Forensic Settings: Theoretical Model and Recommendations for Best Clinical Practice. *International Journal of Forensic Mental Health, 6*(2), 169-183. doi: 10.1080/14999013.2007.10471261

Bernstein, D. P., Nijman, H. L. I., Karos, K., Keulen-de Vos, M., de Vogel, V., & Lucker, T. P. (2012). Schema Therapy for Forensic Patients with Personality Disorders: Design and Preliminary Findings of a Multicenter Randomized Clinical Trial in the Netherlands. *International Journal of Forensic Mental Health, 11*(4), 312-324. doi: 10.1080/14999013.2012.746757

Black, D. W., Blum, N., Eichinger, L., McCormick, B., Allen, J., & Sieleni, B. (2008). STEPPS: Systems Training for Emotional Predictability and Problem Solving in women offenders with borderline personality disorder in prison—A pilot study. *CNS spectrums, 13*(10), 881-886.

Black, D. W., Blum, N., McCormick, B., & Allen, J. (2013). Systems training for emotional predictability and problem solving (STEPPS) group treatment for offenders with borderline personality disorder. *The Journal of Nervous and Mental Disease, 201*(2), 124-129.

Black, D. W., Gunter, T., Allen, J., Blum, N., Arndt, S., Wenman, G., & Sieleni, B. (2007). Borderline personality disorder in male and female offenders newly committed to prison. *Comprehensive Psychiatry, 48*(5), 400-405. doi: 10.1016/j.comppsych.2007.04.006

Black, D. W., Simsek-Duran, F., Blum, N., McCormick, B., & Allen, J. (2016). Do people with borderline personality disorder complicated by antisocial personality disorder benefit from the STEPPS treatment program? *Personality and Mental Health, 10*(3), 205–215. doi: 10.1002/pmh.1326

Blackburn, R., & Coid, J. W. (1999). Empirical clusters of DSM-III personality disorders in violent offenders. *Journal of Personality Disorders, 13*(1), 18-34.

Blacker, J., Watson, A., & Beech, A. R. (2008). A combined drama-based and CBT approach to working with self-reported anger aggression. *Criminal Behaviour and Mental Health, 18*(2), 129-137. doi: 10.1002/cbm.686

Bloom, J. M., Woodward, E. N., Susmaras, T., & Pantalone, D. W. (2012). Use of dialectical behavior therapy in inpatient treatment of borderline personality

disorder: a systematic review. *Psychiatric Services, 63*(9), 881-888. doi: 10.1176/appi.ps.201100311

Blum, N., Pfohl, B., John, D. S., Monahan, P., & Black, D. W. (2002). STEPPS: A cognitive-behavioral systems-based group treatment for outpatients with borderline personality disorder—a preliminary report. *Comprehensive Psychiatry, 43*(4), 301-310. doi: 10.1053/comp.2002.33497

Blum, N., St John, D., Pfohl, B., Stuart, S., McCormick, B., Allen, J., . . . Black, D. W. (2008). Systems Training for Emotional Predictability and Problem Solving (STEPPS) for outpatients with borderline personality disorder: a randomized controlled trial and 1-year follow-up. *The American Journal of Psychiatry, 165*(4), 468. doi: 10.1176/appi.ajp.2007.07071079

Bonnington, O., & Rose, D. (2014). Exploring stigmatisation among people diagnosed with either bipolar disorder or borderline personality disorder: A critical realist analysis. *Social Science & Medicine*, *123*, 7-17.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J. & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40), 2-40.

Dallimore, S. & Griffith, E. (2015). *Involving the wider system in skills-based treatments for Borderline Personality Disorder: A systematic review* (Doctorate in Clinical Psychology thesis, University of Bath, UK). Retrieved from https://researchportal.bath.ac.uk/

Dane, A.V. & Schneider B. H. (1988). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23-45.

Doyle, M., Tarrier, N., Shaw, J., Dunn, G., & Dolan, M. (2016). Exploratory trial of schema-focussed therapy in a forensic personality disordered population. *Journal of Forensic Psychiatry & Psychology, 27*(2), 232-247. doi: 10.1080/14789949.2015.1107119

Evershed, S., Tennant, A., Boomer, D., Rees, A., Barkham, M., & Watson, A. (2003). Practice- based outcomes of dialectical behaviour therapy (DBT) targeting anger and violence, with male forensic patients: a pragmatic and non-contemporaneous comparison. *Criminal Behaviour and Mental Health, 13*(3), 198-213. doi: 10.1002/cbm.542

Fonagy, P., & Bateman, A. W. (2006). Mechanisms of change in mentalization-based treatment of BPD. *Journal of Clinical Psychology, 62*(4), 411-430. doi: 10.1002/jclp.20241

Gee, J., & Reed, S. (2013). The HoST programme: A pilot evaluation of modified dialectical behaviour therapy with female offenders diagnosed with borderline personality disorder. *European Journal of Psychotherapy & Counselling*, 15(3), 233-252.

Gee, J., White, S., Reeves, C., & Bartlett, A. (2016). Forensically-modified Dialectical Behaviour Therapy with Personality Disordered Women Prisoners: Evaluation of Outcomes. *Manuscript submitted for publication*.

Gibbon, S., Duggan, C., Stoffers, J., Huband, N., Völlm, B. A., Ferriter, M., & Lieb, K. (2010). Psychological interventions for antisocial personality disorder. *Cochrane Database of Systematic Reviews (Online), 6*, CD007668.

Hancock-Johnson, E., Griffiths, C., & Picchioni, M. (2017). A focused systematic review of pharmacological treatment for borderline personality disorder. *CNS Drugs*, *31*(5), 345-356.

Herpertz, S. C., Mancke, F., & Bertsch, K. (2016). Aggressiveness, emotional instability and impulsiveness in borderline personality disorder. *Forensische Psychiatrie, Psychologie, Kriminologie*, *10*(3), 189-197.

Higgins, J. P. T., Altman, D. G. & Sterne, J. A. C. E. (2011). Chapter 8: Assessing risk of bias in included studies In: J. P. T. Higgins & S. E. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* (updated March 2011): The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org.

Higgins, J. P. T. & Green, S. (Eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. Retrieved from www.cochrane-handbook.org.

Jamieson, L., & Taylor, P. J. (2004). A re-conviction study of special (high security) hospital patients. *British Journal of Criminology, 44*(5), 783-802.

Johnstone, L. & Boyle, M. with Cromby, J., Dillon, J., Harper, D., Kinderman, P., Longden, E., Pilgrim, D. & Read, J. (2018). *The Power Threat Meaning Framework: Towards the identification of patterns in emotional distress, unusual experiences and troubled or troubling behaviour, as an alternative to functional psychiatric diagnosis*. Leicester: British Psychological Society.

Linehan, M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: New York: Guilford Press.

Linehan, M., Comtois, K. A., Murray, A. M., Brown, M. Z., Gallop, R. J., Heard, H. L., . . . Lindenboim, N. (2006). Two- year randomized controlled trial and follow- up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of General Psychiatry, 63*(7), 757.

Low, G., Jones, D., Duggan, C., Power, M., & MacLeod, A. (2001). The treatment of deliberate self-harm in borderline personality disorder using dialectical behaviour therapy: A pilot study in a high security hospital. *Behavioural and Cognitive Psychotherapy, 29*(01), 85-92.

McCann, R. A., Ball, E. M., & Ivanoff, A. (2000). DBT with an inpatient forensic population: The CMHIP forensic model. *Cognitive and Behavioral Practice, 7*(4), 447-456.

McMurran, M. (2002). Personality Disorder: Expert Paper. *National Research and Development Programme on Forensic Mental Health*.

Mercer, D. E., & Woody, G. E. (1999). Individual drug counselling. *Rockville, MD: National Institute on Drug Abuse*.

Mezzich, A. C., Giancola, P. R., Tarter, R. E., Lu, S., Parks, S. M., & Barrett, C. M. (1997). Violence, suicidality, and alcohol/drug use involvement in adolescent females with a psychoactive substance use disorder and controls. *Alcoholism: Clinical and Experimental Research, 21*(7), 1300-1307.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ, 339*. doi: 10.1136/bmj.b2535

Moore, K. E., Gobin, R. L., McCauley, H. L., Kao, C. W., Anthony, S. M., Kubiak, S., ... & Johnson, J. E. (2018). The relation of borderline personality disorder to aggression, victimization, and institutional misconduct among prisoners. *Comprehensive Psychiatry*, *84*, 15-21.

Moore, K. E., Tull, M. T., & Gratz, K. L. (2017). Borderline personality disorder symptoms and criminal justice system involvement: The roles of emotion-driven difficulties controlling impulsive behaviors and physical aggression. *Comprehensive Psychiatry*, *76*, 26-35.

Nee, C., & Farman, S. (2005). Female prisoners with borderline personality disorder: Some promising treatment developments. *Criminal Behaviour and Mental Health, 15*(1), 2-16. doi: 10.1002/cbm.33

Nee, C., & Farman, S. (2008). Treatment of borderline personality disorder in prisons: Findings from the two dialectical behaviour therapy pilots in the UK. In J. Hagen, & E. Jensen (Eds.), *Personality disorders: new research* (pp. 107-121). New York, NY: Nova Science Publishers, Inc.

NICE. (2009). Borderline Personality Disorder: recognition and management. Manchester, UK: National Institute for Health and Care Excellence.

NIMH. (2003). *Personality disorder: No longer a diagnosis of exclusion. Policy implementation guidance for the development of services for people with personality disorder*. London: National Institute for Mental Health in England. Retrieved from http://personalitydisorder.org.uk/wp-content/uploads/2015/04/PD-No-longer-a-diagnosis-of-exclusion.pdf.

Oldham, J. M. (2015). The alternative DSM-5 model for personality disorders. *World Psychiatry*, 14(2), 234.

Pfohl, B., Blum, N., John, D. S., McCormick, B., Allen, J., & Black, D. W. (2009). Reliability and validity of the Borderline Evaluation of Severity Over Time (BEST): A self-rated scale to measure severity and change in persons with borderline personality disorder. *Journal of Personality Disorders*, *23*(3), 281.

Polaschek, D. L. L., & Daly, T. E. (2013). Treatment and psychopathy in forensic settings. *Aggression and Violent Behavior, 18*(5), 592-603. doi: 10.1016/j.avb.2013.06.003

Raine, A. (1993). Features of borderline personality and violence. *Journal of Clinical Psychology, 49*, 277-281. doi: 10.1002/1097-4679(199303)49:2<277::AID-JCLP2270490222>3.0.CO;2-C

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., & Wells, G. A. (2011). Chapter 13: Including non-randomized studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)*: The Cochrane Collaboration.

Reiss, D., Quayle, M., Brett, T., & Meux, C. (1996). Dramatherapy for mentally disordered offenders: Changes in levels of anger (Vol. 11, pp. 395s-396s).

Salekin, R. T., Worley, C., & Grimes, R. D. (2010). Treatment of psychopathy: A review and brief introduction to the mental model mpproach for psychopathy. *Behavioral Sciences & the Law, 28*(2), 235-266. doi: 10.1002/bsl.928

Samaritans. (2016). Suicide: Facts and Figures. Retrieved August 1, 2016, from http://www.samaritans.org/about-us/our-research/facts-and-figures-about-suicide

Santisteban, D. A., Mena, M. P., Muir, J., McCabe, B. E., Abalo, C., & Cummings, A. M. (2015). The Efficacy of Two Adolescent Substance Abuse Treatments and the Impact of Comorbid Depression: Results of a Small Randomized Controlled Trial. *Psychiatric Rehabilitation Journal, 38*(1), 55-64. doi: 10.1037/prj0000106

Santisteban, D. A., Muir, J. A., Mena, M. P., & Mitrani, V. B. (2003). Integrative Borderline Adolescent Family Therapy: Meeting the Challenges of Treating Adolescents With Borderline Personality Disorder. *Psychotherapy: Theory, Research, Practice, Training, 40*(4), 251-264. doi: 10.1037/0033-3204.40.4.251

Shaw, C., & Proctor, G. (2005). I. Women at the margins: A critique of the diagnosis of borderline personality disorder. *Feminism & Psychology*, *15*(4), 483-490.

Snyder, J. F. (1997). Origins of antisocial behavior: Negative reinforcement and affect dysregulation of behavior as socialization mechanisms in family interaction. *Behavior Modification, 21*, 187-216. doi: 10.1177/01454455970212004

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry, 35*(6), 773-782.

Stoffers, J. M., Vollm, B. A., Rucker, G., Timmer, A., Huband, N., & Lieb, K. (2012). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews, 256*(8). doi:10.1002/14651858.CD005652.pub2

Tarrier, N., Dolan, M., Doyle, M., Dunn, G., Shaw, J., & Blackburn, R. (2010). Exploratory randomised control trial of schema modal therapy in the personality disorder service at Ashworth Hospital. *Ministry of Justice Research Series* (Vol. 5).

van den Bosch, L. M., Hysaj, M., & Jacobs, P. (2012). DBT in an outpatient forensic setting. *International Journal of Law and Psychiatry, 35*(4), 311-316. doi: 10.1016/j.ijlp.2012.04.009

van den Broek, E., Keulen-de Vos, M., & Bernstein, D. P. (2011). Arts therapies and Schema Focused therapy: A pilot study. *Arts in Psychotherapy, 38*(5), 325-332. doi: 10.1016/j.aip.2011.09.005

Warren, J. I., Burnette, M., South, S. C., Chauhan, P., Bale, R., & Friend, R. (2002). Personality disorders and violence among female prison inmates. *The Journal of the American Academy of Psychiatry and the Law, 30*(4), 502.

Warner, S., & Wilkins, T. (2003). *Diagnosing distress and reproducing disorder: Women, child sexual abuse and borderline personality disorder. New feminist stories about women and child sexual abuse: Sexual scripts and dangerous dialogues.* London: Taylor & Francis, 167-186.

Wilson, H. (2014). Can Antisocial Personality Disorder Be Treated? A Meta-Analysis Examining the Effectiveness of Treatment in Reducing Recidivism for Individuals Diagnosed with ASPD. *International Journal of Forensic Mental Health, 13*(1), 36-46. doi: 10.1080/14999013.2014.890682

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2003). *Schema therapy: A practitioner's guide*: Guilford Press.

Zanarini, M. C., Frankenburg, F. R., Reich, D. B., Silk, K. R., Hudson, J. I., & McSweeney, L. B. (2007). The subsyndromal phenomenology of borderline personality disorder: a 10- year follow- up study. *The American Journal of Psychiatry, 164*(6), 929.

Appendix 1

Search Strategy

Category	Search Terms
Context/Setting (anywhere)	(“Correctional” OR “Correctional Institution*” OR “Crime*” OR “Criminal Behavio*” OR “Criminal Conviction” OR “Criminal Justice” OR “Criminal Rehabilitation” OR “Criminal*” OR “Delinquency” OR “Female Delinquen*” OR “Forensic” OR “Forensic Psychiatry” OR “Forensic Psychology” OR “Insanity Defense” OR “insanity defence” OR “insane automatism” OR “Juvenile Delinquen*” OR “Juvenile Justice” OR “Low secure” OR “Male Delinquen*” OR “Medium secure” OR “Mentally Ill Offender*” OR “Offender*” OR “Parole” OR “Perpetrator*” OR “Prison*” OR “Prison nursing” OR “Probation” OR “Probation system”) *AND*
Diagnosis (in abstract or abstact/tite)	(“Borderline Personality Disorder” OR “BPD” OR “emotionally unstable personality disorder” OR “EUPD” OR “Borderline state” OR “Emerging Personality Disorder” OR “Personality Disorder*”) *AND*

Intervention (anywhere)	(“Adolescent Psychotherapy” OR “Behavior Modification” OR “Behaviour Modification” OR “Behavior Therapy” OR “Behaviour Therapy” OR “Brief Psychotherapy” OR “Brief Relational Therapy” OR “Cognitive analytic therapy” OR “Cognitive Behavior Therapy” OR “Cognitive Behavioral Therapy” OR “Cognitive Behaviour Therapy” OR “Cognitive Behavioural Therapy” OR “Cognitive Therapy” OR “CBT” OR “Cognitive behavioral stress management” OR “Cognitive behavioural stress management” OR “Control Group*” OR “Delinquent rehabilitation” OR “Dialectical Behavior Therapy” OR “Dialectical Behaviour Therapy” OR “DBT” OR “DBT-CM” OR “Emotion Focussed Therapy” OR “Emotion Focused Therapy” OR “Emotionally focused therapy” OR “Evidence Based Practice” OR “Experimental Design” OR “Family therapy” OR “Group Intervention” OR “Group Psychotherapy” OR “Group therapy” OR “Individual Psychotherapy” OR “Integrative Psychotherapy” OR “Interpersonal Psychotherapy” OR “Intervention” OR “Intervention study” OR “Mentalization” OR “Mentalization based therapy” OR “Mentalisation based therapy” OR “Mindfulness” OR “Outpatient treatment” OR “Psychiatric Rehabilitation” OR “Psychodynamic Psychotherapy” OR “Psychodynamic*” OR “Psychosocial rehabilitation” OR “Psychotherapeutic Processes” OR “Psychotherapeutic Technique*” OR “Psychotherapy” OR “Psychotherapy, Group” OR “Randomized controlled trial” OR “Randomised controlled trial” OR “Random Sampling” OR “Rehabilitation” OR “Rational-Emotive Psychotherapy” OR “Research Design” OR “Schema Therapy” OR “Schema Modal Therapy” OR “Schema-focussed therapy” OR “Schema-focused therapy” OR “Service evaluation” OR “Social rehabilitation” OR “Systemic psychotherapy” OR “STEPPS” OR “Systems Training for Emotional Predictability and Problem Solving” OR “Therapeutic Community” OR “Therapeutic group*” OR “Therapy” OR “Therapeutic*” OR “Transference focused psychotherapy” OR “Treatment” OR “Treatment Effectiveness Evaluation”)

Appendix 2

Risk of Bias Analysis: Intervention Integrity Criteria

Adherence

Dane and Schneider (1988), cited by Higgins and Green (2011) in the Cochrane Handbook, describe adherence as the extent to which specified intervention components were delivered as prescribed.

Criteria for a judgement of 'Low risk' of bias.	The investigators describe an objective method of assessing adherence and there is reason to believe that adherence is high, as the implemented intervention adheres to the content, frequency, duration and coverage prescribed by its designers (Carroll et al., 2007), Dane & Schneider (1988) suggest that such an assessment will often involve trained observers to supply evaluations of adherence.
Criteria for a judgement of 'High risk' of bias.	An objective method of assessing adherence reveals that adherence was inadequate, or it is stated that there was no method of assessing adherence.

Criteria for a judgement of ‘Unclear risk’ of bias.	Insufficient information about adherence assessment to permit judgement of ‘Low risk’ or ‘High risk’. Alternatively, there was an attempt to measure adherence, but this was not conducted using an objective method, or adherence was assessed but the outcome of the assessment was not reported.

Exposure/Attention

Criteria for a judgement of ‘Low risk’ of bias.	Equal attention, i.e., number, length and frequency of implementation of intervention components (see Dane and Schneider, 1988), must be paid to each group, or analyses are conducted controlling for number of treatment contacts, determining that increased attention did not affect the outcome.

Criteria for a judgement of 'High risk' of bias.	Unequal attention, i.e., number, length and frequency of implementation of intervention components (see Dane and Schneider, 1988), paid to each group, with no analyses conducted to control for any discrepancy in attention received by each group.
Criteria for a judgement of 'Unclear risk' of bias.	Insufficient information about attention to permit judgement of 'Low risk' or 'High risk', or this was not considered/addressed.

Program differentiation

Criteria for a judgement of 'Low risk' of bias.	Safeguards were employed to ensure that the participants in each experimental group received only the planned intervention (Dane and Schneider, 1988) (i.e. no other psycho-social interventions received), or, if participants did receive other interventions, the extent of this was measured or controlled for.

Criteria for a judgement of 'High risk' of bias.	Participants in the study continued to receive other interventions during the study period and the extent of this was not measured or controlled for.
Criteria for a judgement of 'Unclear risk' of bias.	Insufficient information about program differentiation to permit judgement of 'Low risk' or 'High risk', or this was not considered/addressed.

Quality of Delivery – Allegiance Bias

The attitude of the researchers delivering the intervention may also influence the response of those receiving the intervention. If the researchers are not committed to an intervention, or are too committed to an intervention, then the responsiveness of individuals may be affected (Carroll et al., 2007), which could affect outcomes.

Criteria for a judgement of 'Low risk' of bias.	None of the study authors are known to have developed the treatment under investigation, or investigators have considered the implications of clinician enthusiasm towards each of the treatment groups.

Criteria for a judgement of 'High risk' of bias.	Treatment/s used in study have been developed by one or more of the main investigators, and this has not been considered or any attempt made to mitigate against the effect of this. No consideration of implementer enthusiasm.
Criteria for a judgement of 'Unclear risk' of bias.	Insufficient information about allegiance to permit judgement of 'Low risk' or 'High risk', or this was not considered/addressed.

Participant Responsiveness

Measures how far participants respond to, or are engaged by, an intervention (Carroll et al., 2007).

Criteria for a judgement of 'Low risk' of bias.	Investigators have formally measured participant response to the intervention, which may include indicators such as levels of participation and enthusiasm (Dane and Schneider, 1988) and these were found to be high in both treatment and control groups.

Criteria for a judgement of 'High risk' of bias.	Investigators have formally measured participant response to the intervention, which may include indicators such as levels of participation and enthusiasm (Dane and Schneider, 1988) and these were found to be low in the treatment group, the control group, or both.
Criteria for a judgement of 'Unclear risk' of bias.	Insufficient information about participation to permit judgement of 'Low risk' or 'High risk', or this was not considered/addressed.

Tables

See separate file

Figure captions

Figure 1. Flow Chart for the selection of eligible studies.

Figure 2a. Risk of bias summary for controlled studies.

Figure 2b. Risk of bias summary for uncontrolled studies.

Figure 2c. Risk of bias summary across controlled studies.

Figure 2d. Risk of bias summary across uncontrolled studies.